EUROPEANA WHITE PAPER 1

Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana

Professor Stefan Gradmann
Berlin School of Library and Information Science / Humboldt Universität zu Berlin

europeana
think culture

April 2010

Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana

Stefan Gradmann, Berlin School of Library and Information Science / Humboldt Universität zu Berlin. Stefan.gradmann@ibi.hu-berlin.de

1 Europeana: for Whom and to What End?

„Europeana.eu is about ideas and inspiration. It links you to 6 million digital items.“ This is the opening statement taken from the Europeana WWW-site (http://www.europeana.eu/portal/aboutus.html), and it clearly is concerned with the mission of Europeana – without, however, being over-explicit as to the precise nature of that mission.

Europeana's current logo, too, has a programmatic aspect: the slogan “Think Culture” clearly again is related to Europeana's mission and at same time seems somewhat closer to the point: 'thinking' culture evokes notions like conceptualisation, reasoning, semantics and the like.

Still, all this remains fragmentary and insufficient to actually clarify the functional scope and mission of Europeana. In fact, the author of the present contribution is convinced that Europeana has too often been described in terms of sheer quantity, as a high volume aggregation of digital representations of cultural heritage objects without sufficiently stressing the functional aspects of this endeavour.

This conviction motivates the present contribution on some of the essential functional aspects of Europeana making clear that such a contribution – even if its author is deeply involved in building Europeana – should not be read as an official statement of the project or of the European Commission (which it is not!) - but as the personal statement from an information science perspective!

From this perspective the opening statement is that Europeana is much more than a machine for mechanical accumulation of object representations but that one of its main characteristics should be to enable the generation of knowledge pertaining to cultural artefacts.

The rest of the paper is about the implications of this initial statement in terms of information science, on the way we technically prepare to implement the necessary data structures and functionality and on the novel functionality Europeana will offer based on these elements and which go well beyond the 'traditional' digital library paradigm.

However, prior to exploring these areas it may be useful to recall the notion of 'knowledge' that forms the basis of this contribution and which in turn is part of the well known continuum reaching from data via information and knowledge to wisdom.

2 Knowledge: a Challenging Concept

> „There are thing[sic!] we know that we know. There are known unknowns. That is to say there are things that we now know we don't know. But there are also unknown unknowns. There are things we don't know we don't know. So when we do the best we can and we pull all this information together, and we then say well that's basically what we see as the situation, that is really only the known knowns and the known unknowns. And each year, we discover a few more of those unknown unknowns.“
>
> Donald Rumsfeld on „analysis on intelligence information“, 6th June 2002
>
> (http://www.defense.gov/transcripts/transcript.aspx?transcriptid=3490)

As illustrated by the above verbal struggles the former US Secretary of Defense had to get hold of 'knowing', the very concept of 'knowledge' seems to be extremely difficult to grasp. Therefore, at least in the knowledge management literature, most attempts to conceptualise knowledge – rather than giving a definition in the proper sense – end up situating knowledge in a well known conceptual hierarchy and which is well summed up in Bates (2005). This so called DIKW-Hierarchy (abbreviating the terms Data, Information, Knowledge, Wisdom) is usually traced back to T. S. Eliot's famous lines

“Where is the Life we have lost in living?

Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?”

(T.S. Eliot, "The Rock", Faber & Faber 1934)

Information and Knowledge Management literature has added a fourth element to this chain, namely *data*, and the succession of the four elements is usually thought of as a continuum, with no clear binary transitions from one stage to the other.

2.1 Data

The continuum starts with *data*, which – in the context of information science - are usually thought of as discrete, atomistic, small portions of 'givens' (which is the etymological root of 'data') that have no inherent structure or necessary relationship between them. Data exist at different levels of aggregation and abstraction: the raw data obtained from measuring, counting or sensor activity are mostly aggregated to a degree where regularities begin to occur and these aggregated data thus have a potential of being transformed into information. Still, even these higher aggregations of data share an elementary characteristic with raw, unaggregated data: they have no meaning in themselves.

In a linguistic metaphor data could be said to be on phonetical level.

2.2 Information

The transformation to *information* happens once patterns can be discerned in these data – and this is when they start being meaningful. At this level, data are organised into patterns providing – in the words of Ackoff (1989) - “answers to "who", "what", "where", and "when" questions”.

In terms of our linguistic metaphor we are now on phonological and lexical level.

2.3 Knowledge

Knowledge, then, is information that has been made part of a specific context and is useful in this context. The contextualisation processes leading to a specific set of information becoming knowledge can be based on social relations (information as part of a group of people's apprehension of the world, information present in the memory of a person) or semantically based (information related to contextual information via shared properties and thus becoming part of a semantic 'class' of information).

On this level of knowledge it becomes possible, as well, to derive new knowledge (or at least new information) from combined existing knowledge: a form of interpolative – albeit very mechanical – reasoning such as the one based on formal logic in artificial intelligence applications.

With knowledge we clearly are on the syntactic level of the linguistic metaphor.

2.4 Wisdom (or rather thinking?)

This is the last stage of the original hierarchy such as it was first conceived by Ackoff (1989) – and by far the most difficult to grasp.[1]

In the summary of their literature review Rowley and Slack (2008) identify the following facets of '*wisdom*':

- is embedded in or exhibited through action;
- involves the sophisticated and sensitive use of knowledge;
- is exhibited through decision making;
- involves the exercise of judgement in complex real-life situations;
- requires consideration of ethical and social considerations and the discernment of right and wrong;
- is an interpersonal phenomenon, requiring exercise of intuition, communication, and trust.

Considering this very complex set of facets of the 'wisdom' notion it may be useful to reduce the complexity and connotative richness of the concept. At least for the purposes of this contribution I will therefore narrow down the semantics of this level and rather use the term '*thinking*' instead to denote the kind of mental activity we cannot

1 The original DIKW hierarchy includes a layer between Knowledge and Wisdom which Ackoff (1989) calls “Understanding”. That layer combines the reasoning faculties I am situating on knowledge level and 'thinking' in a true, original way. I prefer to separate these two activities and prefer to assign them to two different levels of the hierarchy, namely knowledge and wisdom.

(yet) confer to machines. 'Thinking' in the way we mentally generate works of art or complex scientific theorems which are non-deterministic and in this sense substantially different from deterministic reasoning such as in most 'semantic web' approaches.

Thinking evidently would have to be placed on the 'semantic' level of the linguistic metaphor, whereas other aspects of 'wisdom' would probably have to be placed in the 'pragmatic' realm.

A graphical representation of the DIKT part of the continuum as it will be used as conceptual background of this contribution (and which is derived from the one in Syed (1998)) thus could look like in the figure below:

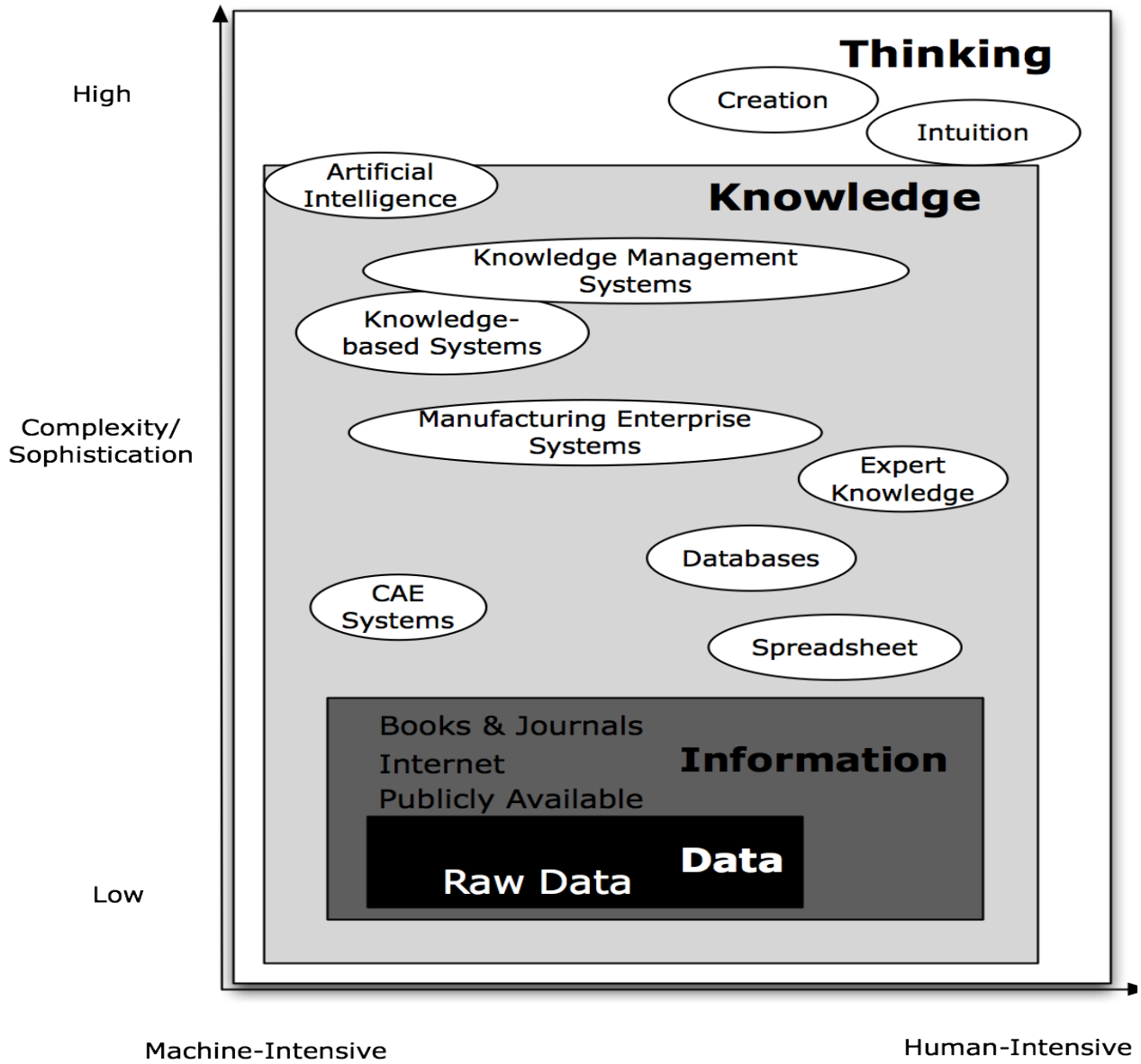

Figure 1: A simplified View of the DIKT-Continuum

3 DIKT in Practice: “Take Five”

Consider the following as a practical illustration of the continuum:

On data level, we perceive an aggregation of pixels such as in the picture below:

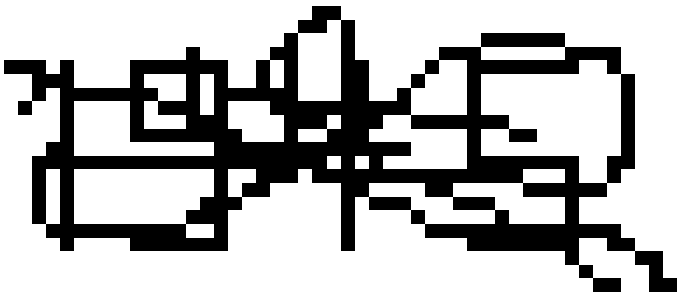

Figure 2: Very Dirty Data

This is a mere aggregation of data with no apparent meaning at all.

However, after removing some of the data noise we are able to identify a pattern in this aggregation which is outlined in the next version of the picture:

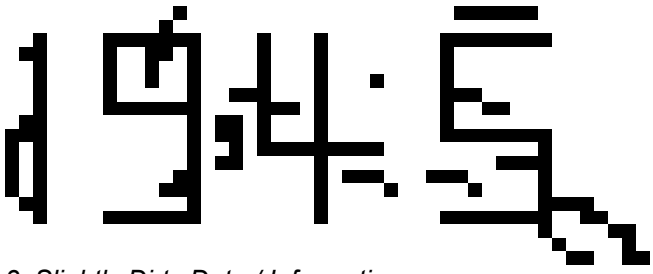

Figure 3: Slightly Dirty Data / Information

- we now are on information level: we have determined a pattern which looks like a sign or a number – and we apply our existing knowledge about 'signs' and 'numbers' to determine the pattern. Note that a machine would probably still have problems identifying the information in this data aggregation! A child without such knowledge about these classes of information objects would not be able to identify the pattern as potentially meaningful, either.

We then move up again one level and consider the cleaned version of the information in semantically formalised context:

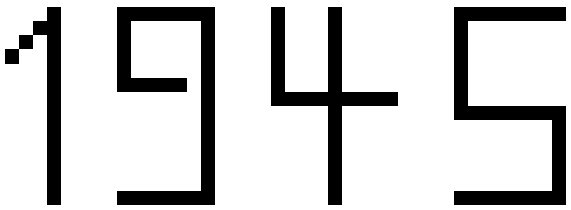

Figure 4: Information with Knowledge Potential

- these are machine detectable string values and a reasoning machine would be able to recognize the string as a number and thus determine the class of unpaired numbers as one relevant context (1945 fitting in the succession 1943 – 1945 – 1947) or even (and this is where we already get close to the upper border of the knowledge level) it may treat '1945' as standing for a year in history and from that infer that the piece of information may belong in the context 'End of World War II'.

One precondition of such reasoning is to embed the reasoning machine in a layer of contextualisation resources such as the rapidly emerging Linked Open Data (LOD) cloud as illustrated in the picture below:

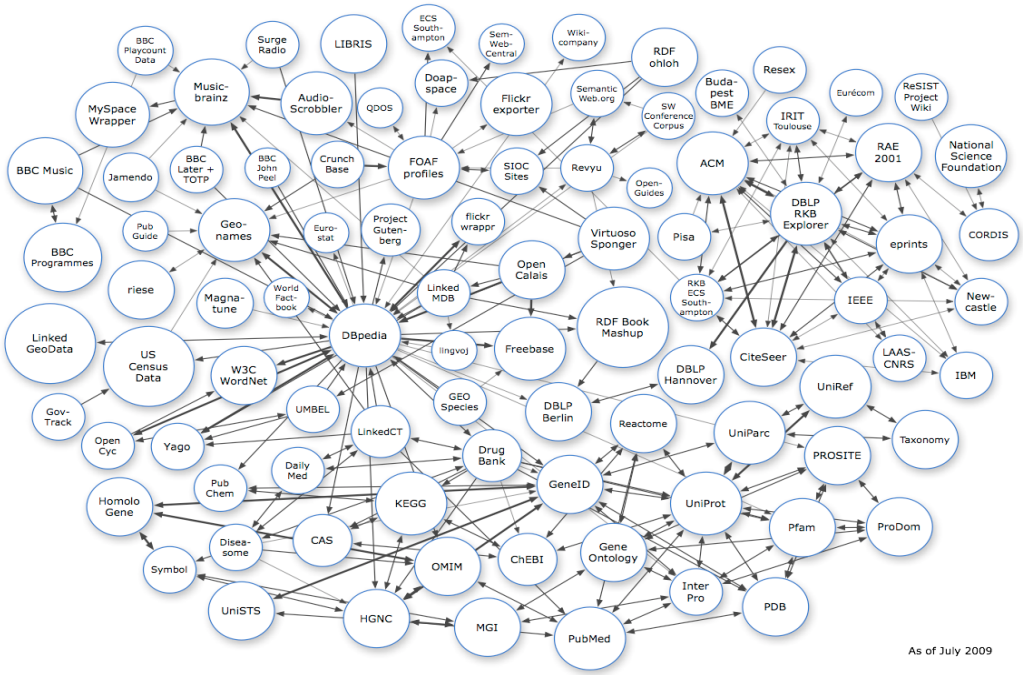

Figure 5: Linked Open Data cloud taken from http://www4.wiwiss.fu-berlin.de/bizer/pub/lod-datasets_2009-07-14.html (created by Chris Bizer)

And finally a human interpreter could consider one digit of the string, the number 5, in isolation and – in the strange ways we as humans 'think' – end up with associating as below

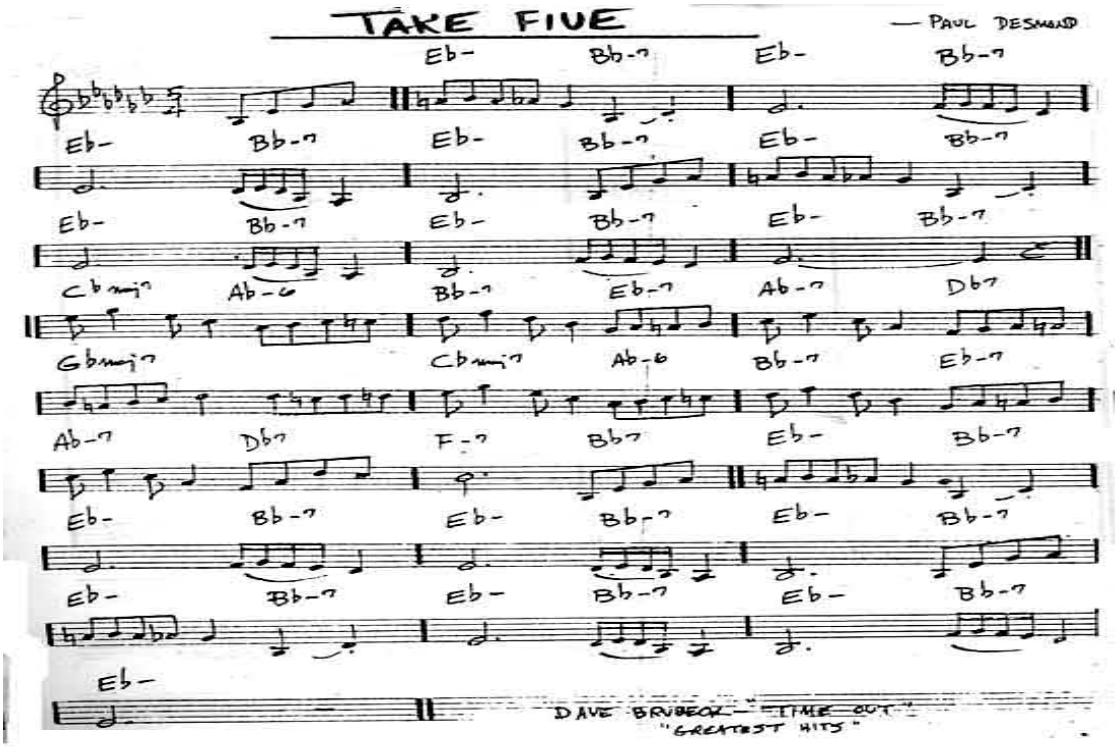

Figure 6: Take 5 Sheet

Or a human might end up humming the tune that goes with this sheet and which is available at http://itunes.om/de/album/dave-brubecks-greatest-hits/id157427923.

Strange as it may seem, this is the way lots of original artwork is conceived and such 'thinking' in terms of mental operations based on shifts of meaning, connotation and personal association context may never fit in any formal model we could conceive.

4 Europeana in the DIKT Continuum

The above recapitulation of the DIKT continuum enables us to return to Europeana and once again consider the mission of this endeavour to bring together millions of representations of cultural artefacts from all kinds of European cultural heritage institutions (and which I refrain from calling a Digital Library for reasons outlined in Concordia, Gradmann & Siebinga (2009))

It should be clear by now that a view of Europeana as a huge agglomeration of data would be terribly inappropriate. However, viewing Europeana as a huge information repository would be almost as inadequate. Instead of such views, we have described the intended characteristics of Europeana as part of what we called a “cultural commonwealth” in the following terms in a recent publication:

> “... we suppose that instead of trying to sustain the digital information silos of the past, cultural heritage communities are ready for an information paradigm of linked data and thus for sharing as much semantic context as possible. Only in such a mental setting does the shift from the portal paradigm to the

> vision of an API as Europea-na's primary incarnation truly make sense.
>
> This mentality shift is a big leap, since it requires cultural heritage institutions to think, not primarily within the boundaries of their particular collections, but in terms of what these collections might add to a bigger, complex and distributed information continuum coupled with various contextual resources enabling European users to turn partial aggregations of this continuum into knowledge that is relevant in their specific context.
>
> The idea thus is not to pre-aggregate information in fixed structures for basically static reuse, but to make it available together with functional primitives for usage scenarios not exclusively defined by Europeana [...]
>
> As part of this mentality shift, cultural heritage institutions will also need to increasingly feel part of a larger community sharing a set of generic standards for organizing information and making it available: the standards referred to here will mostly be created by external instances such as the W3C rather than by the cultural heritage communities themselves!" (Concordia, Gradmann, Siebinga (2009), quoted from manuscript in print)

Europeana should thus be seen as a big aggregation of digital representations of cultural artefacts together with rich contextualisation data and embedded in a Linked Open Data architecture that enables use of these representations in terms of generating knowledge via automated inference operations – or sometimes even as a basis for truly speculative and original thinking in some of the more ambitious scenarios.

The rest of this contribution outlines how we are currently trying to reach this ambitious goal and to which functional end we are doing this work.

5 Semantic Contextualisation in Europeana

In order to understand the following it is important to distinguish the Europeana prototype currently visible at http://www.europeana.eu/portal/ from what is intended to be the result of the two core projects of the Europeana group of projects (more at http://group.europeana.eu/web/guest) The thematic network Europeana Version 1.0 and the project EuropeanaConnect together are working towards implemention of the the functionality and technical characteristics outlined in Dekkers, Gradmann & Meghini (2009). More specifically, WP1 of EuropeanaConnect is working at the creation of the semantic data layer according to the work plan published at http://www.europeanaconnect.eu/workplan.php.

It is important to understand that the metadata currently aggregated and which conform to the Europeana Semantic Elements specification (2009) are not an adequate basis for creating the fully operational Europeana including semantic features as outlined below, and that partial re-delivery of data is a very likely scenario as a consequence. This is part of the overall planning for building Europeana.

A platform much closer to the final goals of the current project phase than the current prototype is available at http://eculture.cs.vu.nl/europeana/session/search. This is a research prototype of a semantic search engine for Europeana created by VU Amsterdam, one of the EuropeanaConnect WP1 partners, and when giving examples at the end of this contribution I am always referring to this research prototype!

5.1 How?

On a very abstract level, Europeana can be seen as a large collection of representations of born digital or digitised cultural heritage objects which themselves remain outside the Europeana data space. In this abstract vision, the representations are linked to each other and additionally are contextualised with links to nodes of a semantic network that forms the second data layer in Europeana. These two links together are used to create rich functionality that is offered on the user interface giving the choice to the user of navigating on either of these levels. This view is illustrated in the figure below

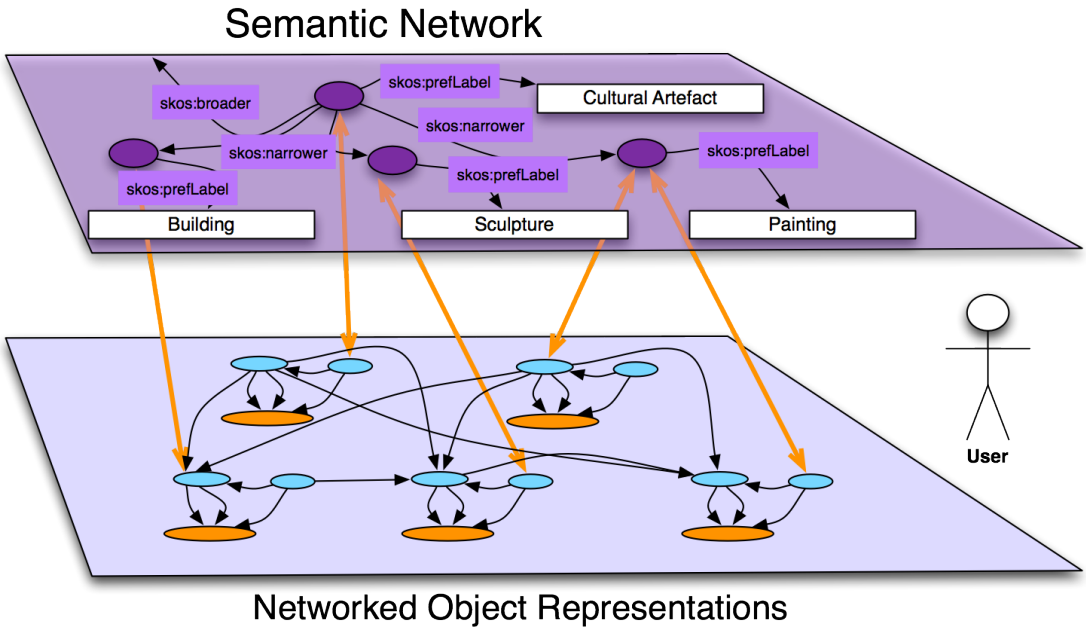

Figure 7: Europeana Data Levels

Furthermore, and as illustrated in Figure 2, these representations (ore:aggregations) are organised as aggregations of web resources in terms of the OAI ORE model representing irw:PhysicalEntityResources within Europeana by means of ore:proxies. Both ore:aggregations and ore:proxies can have contextual links to other aggregations as well as to concept nodes (the circles in purple) such as those representing time and space entities or abstract concepts.

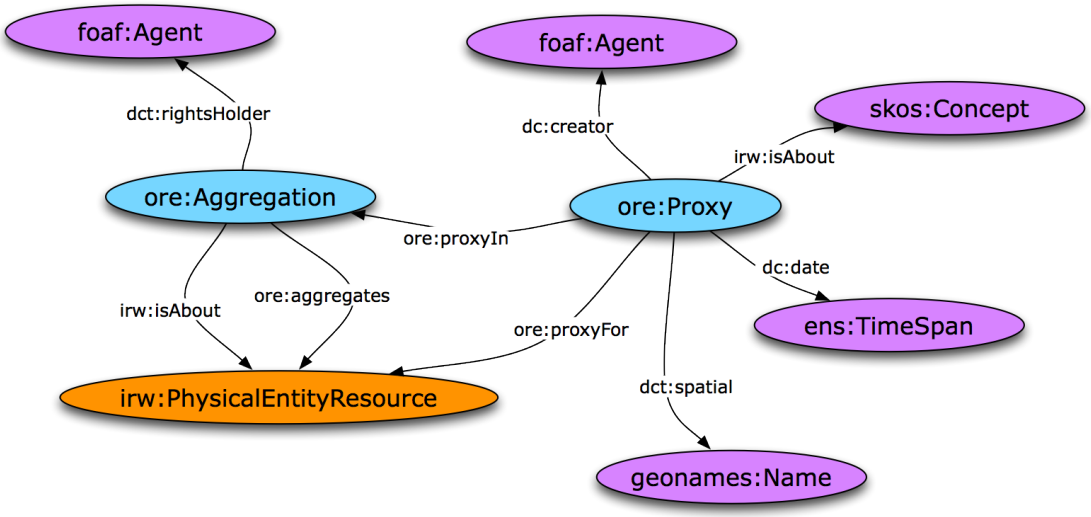

Figure 8: Simplified Europeana Object Representation

Both the internal structure of the object representations and their contextualisation build upon the elements provided by the content suppliers, but substantial parts of this structure and context will be created in the course of the Europeana data ingestion routines.

In terms of a data ingestion and processing workflow for Europeana this implies the following steps.

5.1.1 SKOSification

We assume that in many cases metadata pertaining to digital objects will be provided as records including embedded links to contextualisation resources. These can be links to Linked Open Data (LOD) on the WWW (preferably) or to authority files used within the data supplier's production environment. We also assume that the relevant authority files pertaining to persons, corporate bodies, geographical entities, time periods or other, more abstract concepts are delivered together with the object representation metadata. In such cases we can either reuse the LOD links directly or else we will have to transform the authority file entities into semantic WWW resources expressed in terms of the SKOS standard (and thus having a URI) (cf. Miles & Bechhofer (2009)) and redirect links to these URIs. This process is internally referred to as 'SKOSification'.

Alternatively, and in quite some cases as well, we will not receive pointers to external resources as attribute values but literal terms instead. Such cases have to be dealt with (along with others) in the context of step 5.1.4.

5.1.2 Matching

The semantic contextualisation resources supplied (LOD or authority files delivered) will in many cases be partly redundant with different data suppliers remodelling identical persons or concept resources several times in their respective working environments. Such cases have to be detected systematically in order to (ideally) pull together all entities pertaining to a given concept resource.

5.1.3 Mapping / Merging

Based on such matching operations resources pertaining to one given concept can subsequently either be merged (in case we control all of the resources to be processed in such a way), this results in a new SKOS entity with one preferred term; links to the former (now merged) SKOS entities will have to be redirected.

Otherwise (and this will be systematically the case with LOD, which Europeana by definition doesn't control), entity mappings will have to be established and implemented in such a way as to obtain a result that is functionally similar to actually merging the resources.

5.1.4 Automated Contextualisation of Object Representations

Finally, there will be many object metadata that are not or insufficiently contextualised to fit in the functional model of Europeana. These will have to be contextualised by automatic means as much as possible, creating links to existing contextualisation resources. To do so literal attribute values can be used in many cases if these can be successfully mapped to existing skos:prefLabel values. Algorithms based on co-occurence with other, well contextualised items will be helpful, as well.

The aim is to create a relatively homogeneous semantic context for object representations in Europeana as well as means to automatically position object representations within this context.

5.1.5 Linked Data Integration

The agenda sketched above is already quite complex and ambitious in itself – but gets further complicated and even richer with the massive growth of the so called Linked Open Data environment[2]. Our aim is to integrate the data layer providing semantic context for Europeana object representations as seamlessly in the LOD architecture as possible.

This implies giving up some autonomy: the very idea of 'control' becomes obsolete to some extent that way – but the gain in functionality and rich context will be considerable and – above all – this step makes Europeana part of a much larger community and in a way simply an integrated part of the WWW, the biggest interoperability framework the world has ever seen. In case technical problems (or problems of scalability!) appear in this context we do not have to solve them on our own but share them with millions of others world wide – which is a reassuring idea given the very limited resources Europeana has to ensure maintain regular operations.

5.2 To What End?

As said before, the 'Thought lab' environment can be used to have at least a glimpse at what will be possible on a much larger scale once the agenda depicted above has been operationalised.

2 The slide set presented by Tim Berners-Lee in February 2009 and which is available at http://www.w3.org/2009/Talks/0204-ted-tbl/#%281%29 provides a good introduction to LOD. The “Introduction to Linked Data” presentation by Tom Heath at http://tomheath.com/slides/2009-02-austin-linkeddata-tutorial.pdf provides a good detailed introduction to the field.

Thought lab is largely based on work done by the Free University of Amsterdam in the MultimediaN project and which is described at length in van Ossenbruggen et al. (2007).

The environment is constituted by object representations from 3 museums (Louvre, Rijksmuseum and RKD) together with their semantic context, some of which is owned by these institutions, some of which licensed (mostly from the Getty Institute) and some of which (like WordNet) is part of the LOD world.

This data set probably is a realistic test case for what the Europeana data environment will look like in the future. The data cloud below visualises Thought lab:

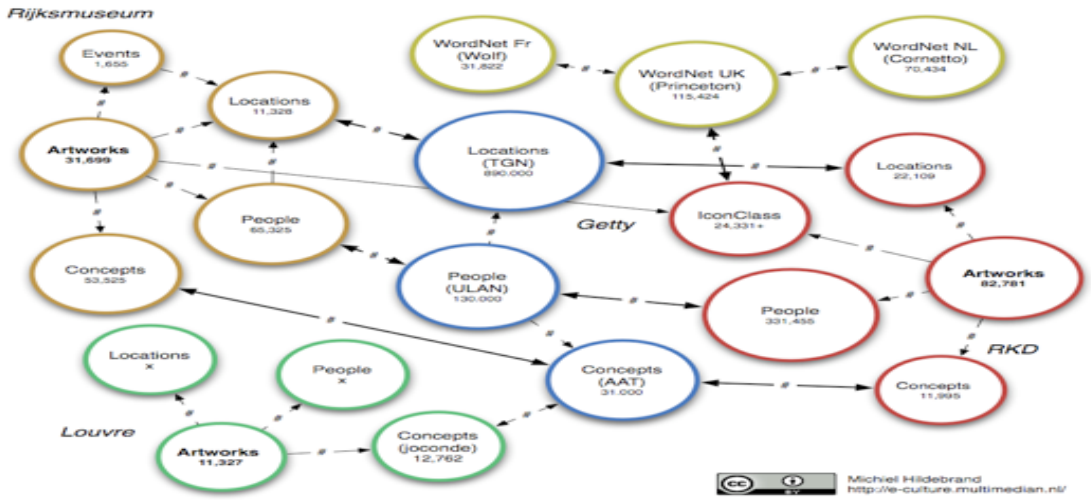

Figure 9: Europeana Thought lab Data Cloud

The architecture of this environment is fully based on W3C standards and more specifically, all information within Thought lab is available as RDF triples. In the example below some of the new functional features enabled are outlined.

This already starts with searching: typing in the search term “Paris” results in dynamic contextual suggestions:

Figure 10: Searching in Thought lab

And once a result set has actually been created more or less surprising items appear in there.

First of all, the system seems to "know" that the Tuileries and the Louvre are located in Paris as is evident from the cluster with the "works showing a more specific location":

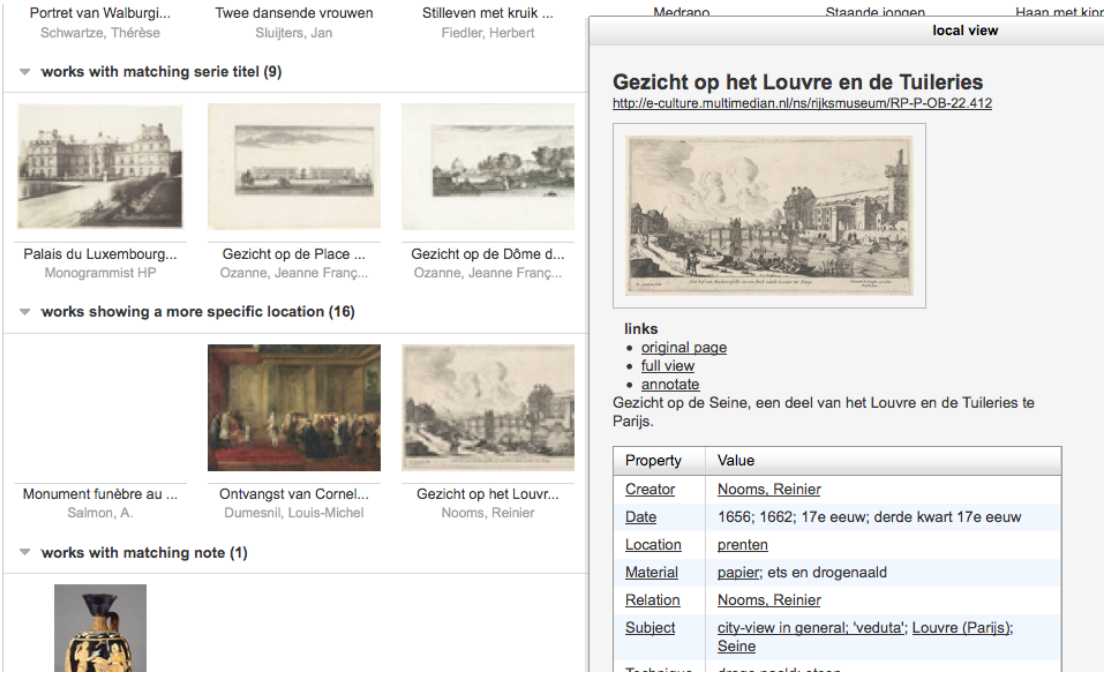

Figure 11: Result Set Details in Thought lab

But – and maybe somewhat more surprising – among the "works showing matching

persons" not only figure four representations of the mythical Paris, but also (as the last one) a painting of the rape of Helena:

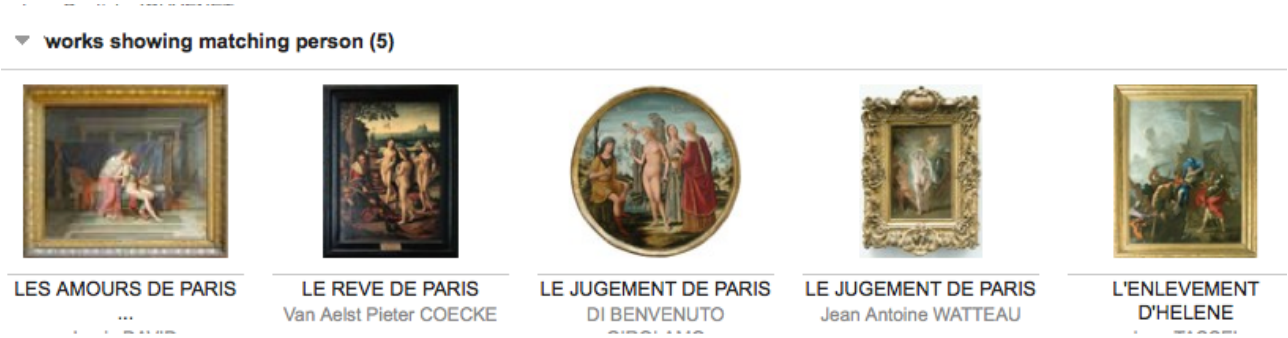

Figure 12: Paris and Helena

However, a look at the attribute set behind shows us that one of the triples (circled in red) is "<painting URI> hasMetadataValue <URI Pâris myth>":

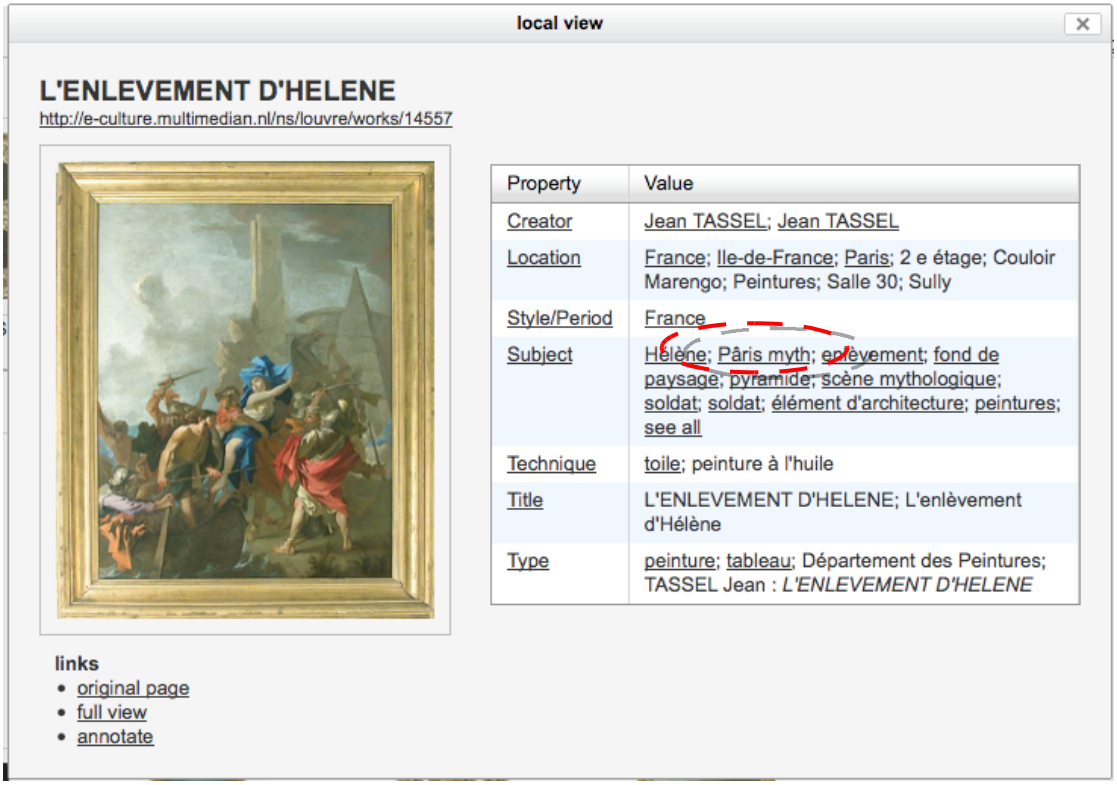

Figure 13: Result Details in Thought lab

- and dereferencing this latter URI takes us to a representation of the Pâris myth with all objects associated in Thought lab:

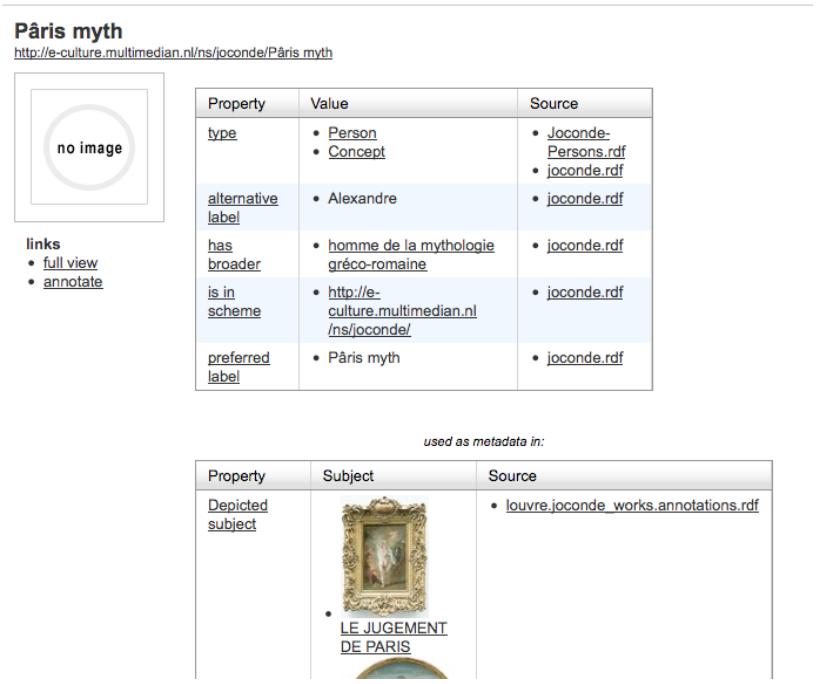

Figure 14: SKOS Node for Paris Myth

And from this rich SKOS node you might be taken to the mythical apple, and from there again to Adam and Eve and into an infinity of triple clusters in Thought lab as well as to newly inferred ones:

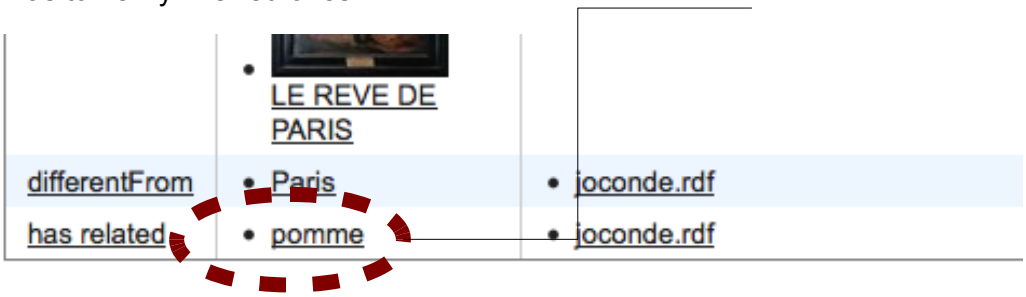

Figure 15: Related Terms

- for it is important to keep in mind that the RDF framework behind this environment can be used both by humans and by machines for very simple reasoning operations based on the RDFS class model.

6 From 'Connecting' to 'Thinking'

This small example should have been sufficient to give an idea of the substantial potential of the approach based on semantic contextualisation which we intend to put to work in Europeana. Once available on large scale such an environment can evolve into a basis for 'Mode 2' knowledge generation frameworks such as discussed in Nowotny, Scott & Gibbons (2003) and Schlögl (2005) or again into semantics based personalised information retrieval environments such as discussed in Vallet (2007) and Vallet et al. (2007).

Actually, the figure below taken from Vallet (2007) bears quite some resemblance with our figure 7 above – and this probably is not by accident!

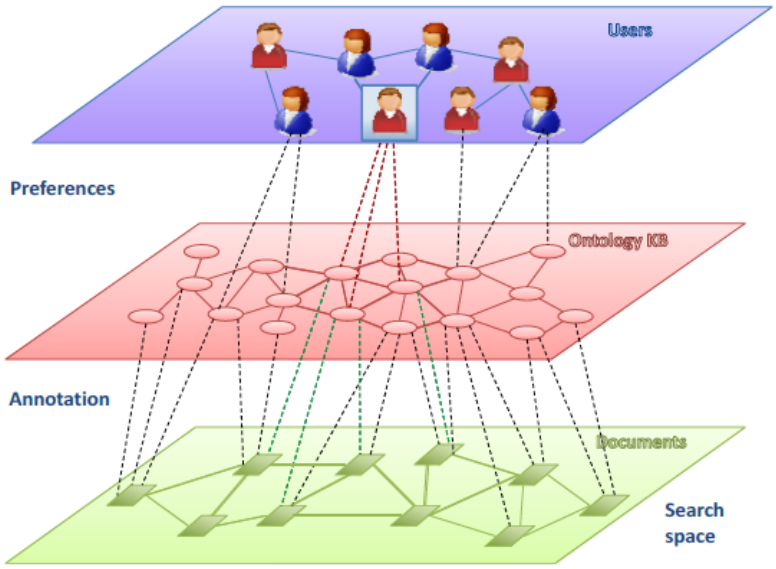

Figure 1. Link between user preferences and search space

Figure 16: Figure taken from Vallet (2007)

These statements lead us back to the beginning of this contribution. It should be clear by now that the environment we are trying to build in Europeana clearly is in the domain of 'knowledge' in the mechanistic (yet very powerful) terms of the semantic web which is all about connecting RDF triples by means of logical operations and typed links – but that it has a potential to also enable creative thinking in a more ambitious sense.

Seen in these terms one perfectly understands why the first logo used for Europeana as shown below has finally been abandoned:

Figure 17: Former Europeana Logo

The keyword here was “connecting” - whereas the keyword in the logo we are currently using for reasons that should be evident from this contribution is “thinking”:

Figure 18: Current Europeana Logo

References

Russell L. Ackoff (1989): From Data to Wisdom. In: Journal of Applies Systems Analysis, Volume 16, pp. 3-9

Marcia J. Bates (2005): Information and knowledge: an evolutionary framework for information science. In: Information Research, Volume 10 No 4 July.

T.H. Davenport, D.W De Long,. M. C. Beers, M. C. (1998). Successful knowledge management projects. Sloan Management Review, 39(2), pp. 43-57

Makx Dekkers, Stefan Gradmann, Carlo Meghini (2009): Europeana Outline Functional SpecificationSpecification Sr development of an operational European Digital Library. Available at http://tinyurl.com/yj4jqpm

Europeana Semantic Elements specifications. Version 3.2.1, 06/11/2009. Available at http://tinyurl.com/yjxnubz

Laurens K. Hessels, Harro van Lente (2008): Re-thinking knowledge production: a literature review and a research agenda. In: Research Policy, vol 37, pp. 740–760

Alistair Miles, Sean Bechhofer (2009): SKOS Simple Knowledge Organization System. Reference. Available at http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Helga Nowotny, Peter Scott, Michael Gibbons (2003): 'Mode 2' Revisited: The New Production of Knowledge. In: Minerva Volume 41, Issue 41, pp. 179-194

Christian Schlögl (2005): Information and knowledge management: dimensions and approaches. In: Information Research, Volume 10 No 4 July

Jaffer R. Syed (1998): An adaptive framework for knowledge work. In: Journal of

Knowledge Management, Volume 2 No 2, December, pp. 59 - 69

David Vallet (2007): Personalized Information Retrieval in Context Using Ontological Knowledge. Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.360&rep=rep1&type=pdf

David Vallet, Pablo Castells, Miriam Fernández, Phivos Mylonas, and Yannis Avrithis (2007): Personalized Content Retrieval in Context Using Ontological Knowledge. In: IEEE Transactions on circuits and systems for video technology, 17, pp. 336 – 344

Jacco van Ossenbruggen, Alia Amin, Lynda Hardman, Michiel Hildebrand, Mark van Assem, Borys Omelayenko, Guus Schreiber, Anna Tordai, Victor de Boer, Bob Wielinga, Jan Wielemaker, Marco de Niet, Jos Taekema, Marie-France van Orsouw, and Annemiek Teesing (2007): Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques. In: Museums and the Web 2007, April 11-14. Available at http://tinyurl.com/ossenbruggen2007

Zeleny, M. (1987): Management Support Systems: Towards Integrated Knowledge Management. In: Human Systems Management, 7(1987)1, pp. 59-70